OTHER RELEVANT RESEARCH - SESSION SUMMARY

Candace S. Culhane

National Security Agency
9800 Savage Rd
STE 6514
Ft. Meade, MD 20755-6514

ABSTRACT

This session provided a forum to present other on-going research of interest to the speech recognition community currently working on the DARPA Hub 4 (Broadcast News) task. The two areas covered were automatic speech recognition of Japanese Broadcast News, and the Hub 5 task - automatic speech recognition of conversational speech over the telephone using multilingual corpora.

1. TOWARD AUTOMATIC RECOGNITION OF JAPANESE BROADCAST NEWS

This presentation was given by Sadoki Furui and provided an insightful update on the state of the art in Japan on recognition of broadcast news. The task in Japan, as it is in DARPA Hub 4, is to do recognition of real broadcast news programs. One of the most interesting revelations was that the level of recognition for phonemes in both English and Japanese are about the same. Neither language is "easier" than the other.

There was an difference in perplexities between the anchor and other speakers in the data results. It was not known what was causing the difference, though it might be due to disfluencies.

Perhaps the most remarkable aspect of the Japanese work is that they do not use any training data other than the speech signal with a transcription, and they are getting results that are comporable to the Hub 4 task community.

2. CONVERSATIONAL SPEECH RECOGNITION

This portion of the session was divided into three presentations:

- Progress and Plans - Candy Culhane
- Technology Perspective - Fred Jelinek
- Feedback and General Discussion - George Doddington, Moderator

2.1 Progress and Plans

The Hub 5 task is the recognition of conversational speech over the telephone in multiple languages. An in-depth description of the task was given at last year's workshop at Arden House {1}. The research life cycle includes a spring evaluation on English corpora, a summer workshop typically hosted by Johns Hopkins University, and a fall evaluation on non English corpora.

The progress in 1996 showed an improvement over 1995 from 49% WER to 39% WER on English (Switchboard) and a general improvement in the non English evaluations, including an improvement from 73% WER to 66% WER in Spanish and a 79% WER to a 75% WER for Mandarin. The first evaluation on Arabic took place in 1996 and had a WER of 75%, which was consistent with other first time evaluations in a non English language.

The 1997 Evaluations consist of an English only task in the spring and non English tasks in the fall. The languages being offered in the fall are Arabic, English, German, Japanese, Mandarin, and Spanish. The test set for the spring eval consists of 20 calls each from the Switchboad 2 and Call Home English corpora. The test sets for the fall evaluation are 20 calls for each language from the non English Call Home corpus. One of the important contributions of the Hub 5 evaluations are the metrics used. These include the use of alternative error rates in addition to the traditional word error rates (WERs), such as lexeme and character error rates.

This year confidence measures will also be required. The speech recognizers are required to generate a confidence measure for each word. This confidence measure is expressed as a probability. NIST then applies a scoring procedure to compute an overall confidence measure based on the individual word confidence measures computed against the result of the recognition. These confidence measures will provide a way to assess how good the confidence measures are.

For sites interested in signing up for the evaluations, participants need to contact Alvin Martin of NIST, at 301-975-3169, or alvin@jaguar.ncsl.nist.gov

Evaluation	Time	Sign Up Deadline
English	Spring 1997	March 1st, 1997
Non-English	Fall 1997	September 1st, 1997

Table 1: Important Dates for Hub 5 Evaluations.

An evaluation specification is available at the web site http://www.itl.nist.gov/div894/894.01/lvcsr.htm

2.2 An LVCSR Technology Perspective

A summary of government sponsored tasks was given along with a history of the SWITCHBOARD corpus. Some new technologies were introduced in speech recognition to be applied against SWITCHBOARD. These included improvements in signal processing, acoustic modeling and language modeling. These also included multi-pass hypothesis search, decision tree analysis, and SRI's WER results on the mode of speech (read, simulated conversation and true conversation).

There are many inadequacies associated with the current techniques. In order to solve the problems in speech research, the community needs to focus on R&D, not evaluations. But the question was asked "How do you balance the research with evaluations?" It is not enough to just let researchers work alone without periodic progress reports that demonstrate measurable results. There was no clear answer; Dr. Jelinek suggested perhaps a panel should look at this issue. It certainly consumes a lot of energy in the community.

Does the community really need a WER? Perhaps not. Is there a use for producing transcripts? Yes.

2.3 Feedback and General Discussion

The remainder of the session consisted of a give and take among all the attendees of the conference. As is often the case at speech recognition workshops and conferences, the hot topics centered on data collection and evaluations.

There are big differences in performance across the various languages and Hub tasks. Are they due to the size of the training data? Call Home is one tenth the size of Switchboard. Could you get 10% WER on Switchboard if the size of the language model was on the order of the Language Model training data for the Wall Street Journal task?

But there are pitfalls in requiring infinite data. An example cited was the field of biology - tons of data had been collected, but it wasn't until the DNA model was created that the field moved forward. The problem in speech recognition is that we are modeling the data, and not the real problem of speech recognition.

Commercial speech companies will always collect more data in order to improve their product. The research community should concentrate on doing more research. It's true that in order to model language in general, people use a lot of data. To model a specific domain (such as Switchboard), it would be better to not use so much data. But perhaps the community needs a lot of data in order to gain the insights necessary to perform well with less data. And on the other hand, large amounts of data can obscure the basic research problem.

It was estimated that if one collected one hundred million words of transcibed speech data, this could consume the entire budget for speech recognition for one year. BBN challenged this estimate, claiming they spend a penny per word, so this would only cost one million dollars.

It was asserted that data should not be an issue; we should have enough already. But how much is enough? One suggestion was that data collection should consume 10% of the budget for speech research, but others suggested it should be as high as 50%.

Some history was provided, reflecting that 2/3rds of the cost of data collection went into the infrastructure (such as corpus development and evaluations). Later, the research community switched to common data collection for multiple tasks, and the infrastructure percentage costs dropped. Therefore, maybe 15% of the speech budget is an appropriate cost. Dr. Jelinek suggested that the price/performance of data collection could also be improved.

Dave Pallet suggested that most places of research are intimidated by large data sets. Also, researchers are testing internally all the time. So why resist evaluations? Mr. Doddington is well known for favoring more frequent evaluations and suggested that it would be nice to reduce the cost (in terms of blood and sweat) per evaluation. This led to a discussion of task stability.

In order to foster good research, the community needs a stable task to work against. But what is the time length of stable? Perhaps it is best to define the stability of the task in terms of accomplishments against it. Using this model, a task would be defined that has an initial WER of 80%, and would be considered completed when a WER of 5% is acheived. Then the task should be redefined to be more difficult, and research begins anew.

It was estimated that a factor of the square root of 2 would be a good reduction in the WER for speech recognition, and that this would require about eight years of good research. A comment was made that even just four years of stability in the task would be welcomed.

A final comment was made that if the acquisition of more data is only used to work against the same old paradigm, then it won’t do any good.

REFERENCES

1. Culhane, C.S., "Session 7 - Conversational and Multi-Lingual Speech Recognition", *Proc. 1996 DARPA Speech Recognition Workshop*, pp. 143-144.